数字人文视域下先秦典籍植物知识挖掘与组织研究*

作者：吴梦成[1] 林立涛[1] 齐月[1] 黄水清[1] 王东波[1] 刘浏[1]
单位：[1]南京农业大学信息管理学院 南京 210095

摘 要：[目的/意义]对先秦典籍中植物进行知识挖掘，构建先秦典籍植物知识图谱，对认识我国古代人民社会和生活状态等具有重要意义。**[方法/过程]**对先秦典籍中植物词进行详尽标注与计量分析。基于 CRF 和多种深度学习模型构建古汉语植物命名实体识别模型，比较分析各模型性能以确定最优模型；设计了面向知识图谱的古汉语植物知识组织模式。**[结果/结论]**基于领域预训练语言模型 SikuRoBERTa 构建的古汉语植物命名实体识别模型性能最优，调和平均值达 85.44%，为基于实体的植物知识挖掘提供了有效方法；构建了先秦典籍植物知识图谱，实现了对先秦典籍中植物实体及其关联知识的聚合与可视化呈现。

关键词：数字人文；先秦典籍；植物命名实体；深度学习；知识图谱
分类号：I206.2 TP391.1

引言

在中华文明发展的历史长河当中，植物作为重要的文学意象和生活资料作用于人类生活的方方面面。“昔我往矣，杨柳依依”中，杨柳被用于传递作者的惜别之情，“蒹葭苍苍，白露为霜”中的蒹葭用于描述可望而不可及的情境。直至现代，植物仍被寄托诸多意象，如仙人掌象征坚强，玫瑰则象征爱情。此外，部分植物因具备重要医学价值，成为许多中草药的原材料，如金银花具有清热解毒之功效，茉莉花可以疏肝解郁等。可见，植物蕴含的远不止其物质属性知识，其背后还蕴含着情感属性、药用价值等关联知识。

目前，面向植物的研究主要有以下几个方面，一是从环境学视角研究植物对大气、土地的影响[1]；二是从生物学视角研究某类植物的功能、性状[2]；三是从中医药学视角研究植物的药用价值[3]；四是从名物学视角研究植物的命名规律及对应客体的渊源流变[4]。

数字人文的兴起为古籍研究带来了新的研究范式，也为古籍中植物知识的挖掘和组织提供了新的方法和视角。我国海量的数字化典籍资源和近年来古汉语信息处理技术的发展为挖掘古籍隐藏的知识提供了有力的数据和技术支撑。在众多古汉语信息处理技术中，命名实体识别是挖掘词语级知识单元的基础步骤，基于语义网技术发展而来的知识图谱则是组织和存储海量知识单元并提供关联知识可视化和检索有效手段。基于命名实体识别和知识图谱技术，对典籍中蕴含的植物知识进行挖掘、组织和呈现，对于发扬中华优秀传统文化，促进古籍中蕴含知识的创造性转化和创新性发展具有重要意义。

本研究选取 25 部先秦典籍为研究对象，详尽标注其中的植物命名实体，并基于 CRF、Bi-LSTM-CRF、和多种深度预训练语言模型进行对比实验探索构建有效的古汉语植物命名实体识别模型，并使用最优模型应用于对《山海经》中植物实体的识别，实现对植物实体的补充。将标注和识别出的植物实体与来自《植物古汉名图考》等外部资源中植物知识进行关联，并构建知识图谱，实现对先秦典籍中植物知识的整理和可视化。

** 本文系国家社科基金重大项目“中国古代典籍跨语言知识库构建及应用研究”（项目编号：21&ZD331）和国家自科基金青年项目“基于深度学习的典籍引书知识图谱构建及应用研究”（项目编号：72004095）的研究成果之一。
作者简介：吴梦成，博士研究生；林立涛，硕士研究生；齐月，硕士研究生；王东波，博士生导师，教授；刘浏，硕士生导师，副教授；通信作者，Email：db.wang@njau.edu.cn。

相关研究

植物知识挖掘

基于古汉语典籍文本的植物知识挖掘，主要集中于从中医药学视角对植物的药用价值进行研究。如邹俐[5]通过主流本草和本草方剂丛书对山豆根的有关信息进行考证，发现古籍中记载的山豆根无毒与现代研究结果不同，认为山豆根具有易混淆的特点。曲保全[6]通过整理中华医典数据库中含生发、乌发等方剂的记录来研究中医典籍中外用美发方剂的用药规律。袁代昌[7]以中华医典和本草古籍为基础，对乌药的名称、形态、产地等进行本草考证，指出乌药在品质、采收、炮制、功效等方面仍存在部分认知差异。

此外，部分学者也从文学视角对植物命名及其意象进行研究。譬如，于娜娜等[8]对《诗经》中的水生植物和湿生植物意象进行剖析，在挖掘其实用价值的同时还探析了其内在的情感文化价值。谭宏姣[9]全面系统的研究了古汉语植物的命名，总结归纳了植物命名的特点和规律。王薇[10]以《仪礼》中的名物词为研究对象，对其中的八个单音节名物词进行源义考求，并区分了其中所包含植物的物类类别。王凌云[11]采用“文本细读”的方法，以“植物意象”所包含的四层意蕴为线索对当代新诗作品中的植物进行分析和研究，并阐释了植物意象的意义生成方式。马开颜等[12]以典型文学作品中的植物词条为研究对象，通过主题模型探索植物、词汇与主题的相关性，挖掘植物意象在文学作品中与在特定表达之间的对应关系。

综合来看，研究者多以单一古籍为研究对象，研究语料整体规模不大，研究方法自动化程度低、结果展示不够直观。

命名实体识别

命名实体识别作为自然语言处理的基础性研究，历经了从基于规则的方法到统计机器学习方法再到基于注意力机制、图神经网络[13]等深度学习方法的演变。随着计算机技术的发展，命名实体识别的技术也在不断更新、优化，应用前景也越发广泛，已成为数字人文研究的重要技术手段之一。

对汉语古籍文本进行命名实体识别的研究已取得丰硕成果，基于机器学习与深度学习的方法是当前的主流方法，其免去特征工程的同时亦能取得优良的识别效果。相关研究如李娜等[14]将数字化的典籍《方志物产》作为语料，对语料标注后，使用 CRF 对语料中的各种名称包括人名、地名、别名、引用名等多类型命名实体实现有效识别。徐晨飞等[15]以《方志物产》云南卷为语料，完成对语料中人物、产地、引书、物产别名等实体的识别，发现 Bi-LSTM-CRF 模型对引书实体的识别能力更优，BERT 模型对人物实体识别效果最优。王菁薇等[16]以《伤寒论》为语料，构建 ALBERT-BiLSTM-CRF 模型完成对其中疾病、处方、药物、症候、症状等命名实体识别的任务，发现此模型相比其他模型效果更佳。此外，部分学者在神经网络结构上进行了创新，如 Y.Wang[17]提出了一个多态图注意网络（PGAT），旨在从多个维度捕获字符与匹配词之间的动态相关性，以增强字符表示。

综合来看， CRF、Bi-LSTM-CRF、BERT 等是现阶段实现命名实体识别的常用模型，研究对象包括方志物产文本、中医药典籍文本等。另一方面，相关研究均采用面向汉语文本处理的通用模型，而古汉语语法、词法、句法、体裁、语体风格与现代汉语差异较大，因此上述研究可能存在模型结构与文本内容契合度不高，识别结果不全面的问题。

近年来，SikuBERT[18]等面向数字人文研究的古汉语预训练语言模型为古文智能信息处理带来了全新的选择。在此背景下，本研究在构建先秦古汉语植物实体语料库的基础上，基于面向古汉语文本智能处理的预训练语言模型构建有效的古汉语植物实体识别模型，并利用所构建的识别模型辅助植物知识图谱的构建。

知识图谱

知识图谱又称为知识领域映射地图，是一种新型的知识表示形式，能够以可视化的方式组织和呈现某个领域的概念、概念属性及不同概念间的语义关系[19]。该技术已被运用到

领域知识建模[20]、自动问答[21]、主题演变分析[22]等众多领域，特别在中医药古籍知识组织方面，知识图谱有广泛的应用。譬如，张君冬[23]以不孕症为例构建知识本体呈现该领域内的概念关系，并采用数据挖掘方法完善本体语义关系，实现了不孕症中医临床试验知识的语义映射和结构化表达。张向先等[24]使用自顶向下方法构建了敦煌吐鲁番医药文献本体模型，并在此基础上构建知识图谱，实现了敦煌吐鲁番医药文献的知识组织与可视化。翟东升等[25]通过深度学习信息联合抽取模型对中医药专利文本中的实体及关系进行抽取，基于中医药知识图谱本体结构完成了知识图谱的构建。羊艳玲等[26]阐述了中医医案知识图谱构建方法，并以医案中的疾病、症状、药物等实体为例进行命名实体识别和抽取，构建知识图谱，探索其中关系。李贺等[27]以简帛医药文献为研究对象，构建了简帛医药书目本体和内容本体，并以此为基础实现了简帛医药文献知识图谱可视化呈现。

目前，知识图谱也成为了数字人文研究的重要方法，如崔竞烽等[28]以古典诗词为研究对象，使用深度学习模型挖掘诗词中的菊花相关知识以及菊花诗词文本关联。张云中等[29]梳理了历史人物数字资源，构建红色历史人物知识图谱，搭建红色历史人物问答平台。刘欢等[30]以《左转》为研究对象，通过 SVM 和 BERT-LSTM-CRF 模型实现问句意图识别和问句实体识别，构建领域知识图谱并基于 Flask 框架完成问答系统平台的搭建。范青等[31]构建了非物质文化遗产知识图谱，形成关联数据，呈现非物质文化遗产隐形关系。钟远薪等[32]针对艺术图像领域，构建艺术图像知识图谱，对比传统数据库，论证了知识图谱在知识组织应用上的先进性。

从上述研究可以看出，知识图谱在知识组织和知识可视化以及关联分析方面具有优势。与此同时，科技发展带来领域知识不断增加，这使得通常的领域知识图谱需要不断更新旧知识和补充新知识，构建和维护成本高。然而，面向数字人文研究的知识图谱构建则是基于有限的历史典籍，这便赋予其极高的稳定性，减少了后期维护的工作。稳定的领域知识图谱能够给知识检索、自动问答等知识图谱应用提供重要保障。因此本研究选择知识图谱对典籍中植物及其关联知识进行组织和存储。

数据集构建及植物词分布特征统计

数据来源

本研究选取南京师范大学构建的先秦典籍语料库作为研究对象。该语料库含 25 部先秦典籍，按照四部分类法可将其分为"经、史、子、集"四个大类[33]，具体如表 1 所示。该语料内容丰富，涵盖古代军事、文化等多个方面，比较全面地揭示了先秦时期古代人民的生活状态和社会风貌，同时也记述了大量植物，具备较高研究价值。

表 1 25 部先秦部典籍及对应四部分类

典籍种类	典籍名称
经部	《诗经》《尚书》《周礼》《仪礼》《礼记》《谷梁传》 《公羊传》《周易》《左传》《论语》《孝经》《孟子》
史部	《国语》
子部	《孙子兵法》 《吴子》《管子》《老子》《荀子》《庄子》 《韩非子》《墨子》 《吕氏春秋》《商君书》《晏子春秋》
集部	《楚辞》

植物词标注

植物命名实体即指代植物的词，下文简称植物词。在 25 部先秦典籍中，存在一些指代较为宽泛的植物词，无法较为明确地对应到具体的植物品种，如"树""藻""水草"

“荒草”，此类植物词不纳入后续的标注与统计范围。标注工作采用人工标注辅以词典匹配的方式完成，包括词典匹配预标注、人工校对与补充标注三个步骤。

首先，本研究通过以下两个数据源构建古汉语植物词词典。一是《尔雅》中的《释草》和《释木》章节，《尔雅》成书时间与先秦典籍相近，其中《释草》和《释木》记述了大量植物相关的内容。二是《植物古汉名图考》，该书是高明乾历时 30 余年对植物古汉名进行考证的重要成果，其中包含植物古汉名 4394 个，记载的古植物种类丰富。本研究邀请三名具有植物学研究背景的研究生通过人工判读的方式识别和整理上述数据源中的植物词，形成植物词集合。随后对植物词集合中的所有植物词归并去重，形成最终用于词典匹配标注的古汉语植物词典。

接着，利用自编 Python 程序，采用最大逆向匹配策略，对语料中词性为名词“n”的内容进行基于古汉语植物词典的预标注。

最后，对词典匹配标注的结果进行人工校对和补充标注，以提高标注的准确性与全面性。本文作者和上述三名具有植物学背景的研究生分组（每组 3 人）完成本部分工作。每组人员先分别对预标注结果进行校对与补充，再分别对另一组的校对和补充结果进行检查和确认。所有人员参照古诗文网(https://www.gushiwen.cn/)提供的古文与白话译文平行语料库完成本部分工作，以提高对典籍内容理解的准确性。

遵循上述步骤，完成标注后一条语料样例如“【黍】/n 、/w 【粱】/n 、/w 【稻】/n 皆/d 二/m 行/n”。

植物词分布特征统计

在 25 部先秦典籍中，共发现 4576 个植物词，不重复植物词个数为 364 个。在标注过程中，并未发现《孝经》中出现植物词。就植物词总数而言，总数最多的是《仪礼》，包含植物词 635 个，其次是《管子》《诗经》和《礼记》分别是 538 个、472 个、457 个；就不重复植物词数量而言，数量最多的是《诗经》，含不重复植物词 134 个，《管子》和《礼记》次之，分别为 111 个和 95 个。在植物词的全文总次数占比方面，占比最高的是《诗经》，约占 1.36%，也是唯一一部植物词占比超过百分之一的典籍。植物词在不同典籍中的总数、不重复词个数、及占全文字数比值如表 2 所示。

表 2 先秦典籍中植物词数量统计

典籍	植物词数	不重复植物词数	全文占比	典籍	植物词数	不重复植物词数	全文占比
仪礼	635	45	0.62%	管子	538	111	0.31%
诗经	472	134	1.36%	礼记	457	95	0.34%
周礼	302	46	0.40%	吕氏春秋	274	84	0.22%
韩非子	246	51	0.19%	楚辞	232	90	0.77%
左传	206	71	0.10	墨子	205	54	0.23%
晏子春秋	184	37	0.28%	国语	150	40	0.17%
荀子	144	52	0.14%	庄子	141	59	0.15%
尚书	78	32	0.22%	孟子	76	20	0.16%
周易	56	17	0.20%	商君书	46	11	0.18%
谷梁传	44	19	0.1%	公羊传	32	13	0.07%
论语	22	14	0.11%	孙子兵法	20	7	0.16%
老子	11	4	0.14	吴子	5	3	0.07%

从表 2 可见，诗赋类典籍包含的植物词数量占比较高，而经史类占比则相对较低。可以推断植物在中国古代人民生活当中用于写诗作赋是一种较为常见的现象。《诗经》在植物词总数排名、不重复植物词个数排名以及植物词全文占比方面都排名靠前，这一结果其实并非偶然。《诗经》中的诗歌多用“赋比兴”的写作手法，“比”是以彼物比此物，“兴”则是先言他物引所咏之词，《诗经》中 “彼物”和“他物”常常是植物词，如《卫风·硕人》中“手如柔荑”用植物柔荑来表现美人手的柔嫩；《蒹葭》中“蒹葭苍苍，白露为霜。”诗人借“蒹”和“葭”两种植物来抒发自己的思念之情。因此，《诗经》中的植物词在种数和频数方面都位居前列。

先秦典籍植物知识图谱构建

植物知识图谱的构建技术路线如图 1 所示，具体为：（1）古汉语植物实体识别模型构建和应用。通过对比分析法将植物语料输入机器学习模型和多种深度学习模型中，构建面向典籍的植物命名实体自动识别模型，比较最终各模型的植物识别效果，并确定最优模型（2）以《山海经》文本为应用对象，使用最优模型去识别其中的植物，并人工核验植物识别结果，为知识图谱扩充数据源。（3）植物知识图谱构建。以 25 部先秦典籍的文本内容和从互联网百科知识库爬取的植物知识和从《山海经》中抽取的相关植物知识为数据源，结合实体链接和知识融合等方法构建先秦典籍植物知识图谱。

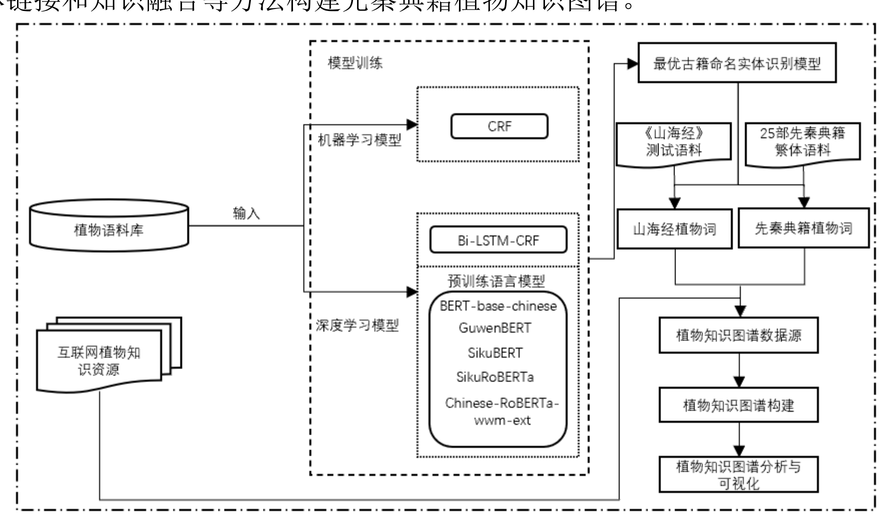


图 1 先秦典籍植物知识图谱构建技术路线

知识表示

知识图谱通过图的形式呈现实体和实体关系，通过三元组的方式组织实体数据和实体关系数据。具体而言，在知识图谱中“节点－边－节点”这样的关系可以看作是“主语－谓语－宾语”的关系，视为知识图谱中的一条记录。具体古汉语植物词及其关联知识的表示如图 2 所示。

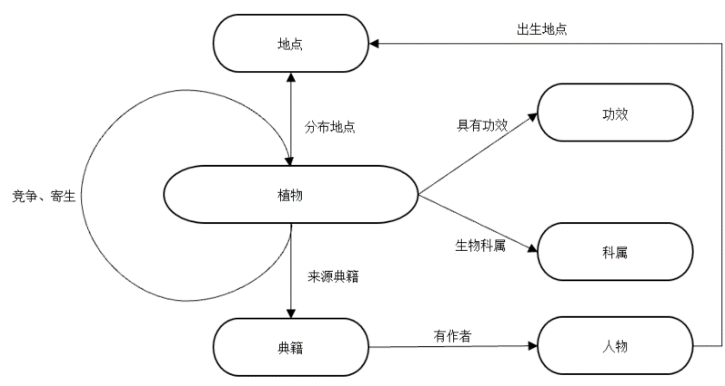


图 2 先秦典籍植物知识表示模型

整个知识图谱皆由此类三元组构成，对于同一主语通常包含多种关系，随着知识不断累积，知识图谱的实体关系网络也会不断扩大，最终知识图谱将会包含海量数据和知识。图 3 展示了一个实例，“植物唐具有功效治疗肾虚腰痛”，植物唐是主语，具有功效是谓语，治疗肾虚腰痛则为宾语。

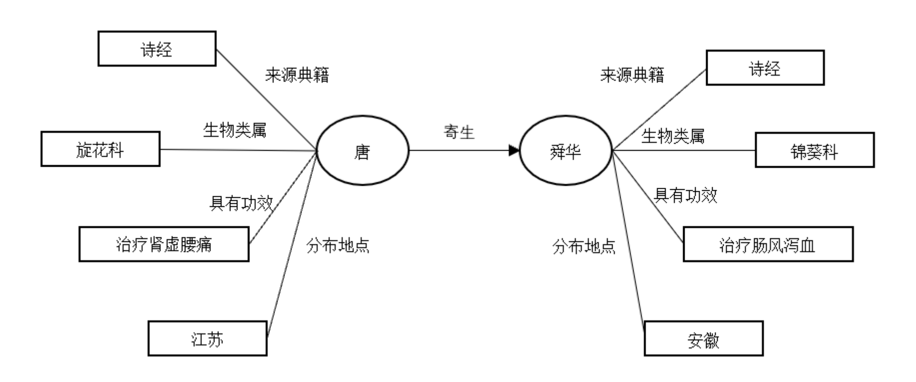


图 3 先秦植物知识表示实例

知识抽取

知识抽取的目的是实现对典籍中植物命名实体的识别和抽取。本研究采用序列标注模型完成知识抽取，采用的具体模型有 CRF、Bi-LSTM-CRF 和多种预训练语言模型，下面简要介绍各类模型。

（1）CRF

条件随机场（Conditional Random Field，CRF）是一种判别式无向图模型。当其应用到标注问题中，就成为根据输入序列对输出序列进行预测的判别模型。其学习方式是在给定训练数据集的基础上通过极大似然估计获得条件概率模型，若进行预测，则是在给定输入序列的基础上寻求条件概率最大的输出序列。

（2）Bi-LSTM-CRF

Bi-LSTM-CRF 由 Bi-LSTM 与 CRF 构成。CRF 层可以通过学习数据集中标签之间的转移概率从而修正 Bi-LSTM 层的输出，提高模型预测准确率。

（3）深度预训练语言模型

深度预训练语言模型是一种基于大规模无监督语料，通过自监督的方式训练而得到的含有语料中词法、句法、上下文信息的语义表示模型。采用领域化的预训练语言模型可以进一步提高其在对应语料上的下游任务性能，因此本研究特别选择面向数字人文的古文预训练模型 SikuBERT、SikuRoBERTa[34]进行实验。SikuBERT 和 SikuRoBERTa 是由南京农业大学基于四库全书语料训练而成，在预训练过程中，二者的词表均使用不含标点符号的繁体中文，且句子切分是以字为粒度。其中 SikuBERT 是基于 BERT-base-Chinese 在《四库全书》语料继续训练得到，在预训练过程中移除了对性能提升帮助不明显的下一句预测任务 SikuRoBERTa 是基于中文版 RoBERTa-Chinese（结合全词遮罩方式）在《四库全书》语料上继续训练得到。用于继续训练的《四库全书》语料为文渊阁版的繁体字《四库全书》正文文本（不含注释），训练数据总字数约 5.3 亿字左右。

为充分比较并筛选出最优模型，本研究还选择了 guwenBERT(https://github.com/Ethan-yt/guwenbert) 、 BERT-base-Chinese[35] 和 Chinese-roberta-wwm-ext[36] 进行对比实验。guwenBERT 是由北京理工大学基于 RoBERTa-Chinese（结合全词遮罩方式）在殆知阁古代文献语料上继续训练得到，其中殆知阁古代文献语料包含 15694 本古籍，总字数约 17 亿字。BERT-base-Chinese 是由谷歌基于中文维基百科数据训练而成，面对中文自然语言处理任务，具有较好的通用性。Chinese-roberta-wwm-ext 是哈工大讯飞联合实验室采用全词掩码技术基于中文通用语料开发的中文预训练语言模型。guwenBERT 是由北京理工大学基于 RoBERTa-Chinese（结合全词遮罩方式）在殆知阁古代文献语料上继续训练得到，其中殆知阁古代文献语料包含 15694 本古籍，总字数约 17 亿字。

语料预处理

在构建植物命名实体自动识别模型前，需要完成对语料的预处理。通过对语料中所有植物词词长的统计，最终确定采用 5 词位标记集作为预处理过程的标注规范。5 词位标记

集可表示为 R={B-P, E-P, M-P, S-P, O}，其中“B-P”表示植物词的起始字符，“M-P”表示植物词的中间字符，“E-P”表示植物词的结束字符，“S-P”表示单字植物词，“O”表示除植物词组成部分以外的所有其他字符。经预处理后的语料样例如表 3。

表 3 古汉语植物词语料预处理结果样例

序号	字符	标记	序号	字符	标记
1	荆	O	6	梓	S-P
2	有	O	7	梗	S-P
3	長	B-P	8	柟	S-P
4	松	E-P	9	豫	B-P
5	文	O	10	樟	E-P

模型构建

本实验所需的计算机配置如下：操作系统为 CentOS 3.10.0，CPU 为 4 颗 Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz，内存大小 256G；GPU 为 6 块 NVIDIA Tesla P40，显存大小 24G。CRF 和 Bi-LSTM-CRF 采用默认训练参数。由于 SikuBERT、SikuRoBERTa 和 BERT-base-Chinese 等深度预训练模型神经网络架构相同，故实验时设置相同训练参数，如表 4 所示。

表 4 深度预训练模型的训练参数

超参数	含义	值
train_batch_size	每次输入模型的句子数	64
max_seq_length	允许输入的最大句子长度	128
epoches	训练轮次	3
leraning_rate	学习率	2E-5
warmup_proportion	预热学习率	0.4

本研究选取精确率（P）、召回率（R）和调和平均数（F1）[37]对模型性能进行评测。表 5 展示了各模型的测试结果，从中可以看出 SikuBERT 与 SikuRoBERTa 的性能较为突出，其中 SikuRoBERTa 模型表现最优，F1 值达 85.44%。

表 5 各植物词自动识别模型评测数值

序号	模型	准确率(P)	召回率(R)	调和平均数(F1)
1	CRF	86.31%	68.87%	76.40%
2	Bi-LSTM-CRF	82.38%	58.53%	67.98%
3	BERT-base-Chinese	80.40%	72.32%	76.15%
4	GuwenBERT	67.19%	53.71%	59.64%
5	SikuBERT	79.62%	83.71%	81.61%
6	SikuRoBERTa	81.54%	89.73%	**85.44%**

模型应用

《山海经》同样成书于先秦时期，其记载了关于古代地理、历史、动物、植物、医学等方面的诸多内容。本研究将上述最优模型应用于《山海经》文本，以实现对古汉语植物词典的补充。最优模型对《山海经》中一句话的识别结果为“其葉如【榆】葉而方，其實如【赤菽】，食之已聋。”，其中“赤菽”一词并非训练语料中的标注实体，该模型能将其准确识别出来，说明本研究构建的模型具备较好的实用效果。最终总计从《山海经》文

本中共识别出中植物词 366 个，121 种，其中既包含在训练语料中已标注的植物词，也新识别出许多植物词如“棪”“榣”等。

知识融合

通过知识抽取虽然获得实体、关系、属性等知识，但是知识来源不同导致抽取到的数据存在很多噪声数据和重复数据，对此类数据的清洗和整合有利于先秦典籍植物知识图谱的完善和优化。知识融合的过程主要分为两步，分别为实体链接和知识合并。

实体链接主要是将数据采集过程中来自不同数据源的重复知识进行融合，即让含义相同的实体合并为一个实体。本研究由于数据源有限，均采用人工判别的方式来判断实体间是否具有相同含义。如网站“植物通”(https://www.zhiwutong.com/)中的“国内分布”和网站“植物智”(https://www.iplant.cn/)中的“分布地”都是对植物在中国境内地点的表述，属于相同实体，应该将二者进行合并。

知识合并主要是在实体链接的基础上，将同一植物实体在不同来源网站上属性进行融合，如“稻”在“植物通”记载了“植物智”中未出现的别名相关知识，“植物智”中又记载了“植物通”中未出现的功用价值、生态习性等知识，将不同来源但是为同一实体的属性内容进行人工选择与合并可以全面吸收植物相关知识，让先秦典籍植物知识图谱内容更加丰富全面。

本研究采用两种方式实现植物知识抽取：一方面先通过人工标注先秦典籍中的植物词，然后利用基于标注数据训练得到的 SikuRoBERTa 模型自动识别获得 25 部先秦典籍和《山海经》中的植物词并进行人工校对，另一方面通过 Python 爬取“植物通”和“植物智”两个网站的植物关联数据，经过数据加工和整合获得最终用于典籍植物知识图谱构建的结构化数据。

在数据呈现方式上，“植物通”和“植物智”多以半结构化形式进行存储；在数据内容上，“植物通”主要存储植物的科名、属名、植物志、别名、来源、性味、功效、国内分布、国外分布、海拔高度、习性、药用部位、药用功能、药用主治、考证、化学成分等属性和关系。“植物智”主要存储植物的学名、俗名、异名、分布地（中文）、形态特征生态习性、图片、标本、标本分布、植物志、保护等级、保护价值、保护措施、栽培要点等属性和关系。将上述数据内容上的知识进行对比整合有利于保证先秦典籍知识图谱中关系的全面性和属性的详尽性。

知识存储

本研究选取图数据库 Neo4j 来构建先秦典籍植物知识图谱。Neo4j 一方面支持使用 Cypher 查询语言实现对实体和实体间关系的语义查询，另一方面在关联度较高的数据上拥有更快的查询速度，且提供了可视化的查询功能。

结合构建的先秦典籍植物命名实体识别模型识别出的植物和从外部数据库爬取的多维植物知识，依照知识表示模型的数据结构，将人工标注先秦典籍中的植物词、利用模型从《山海经》中识别出的植物词和“植物通”和“植物智”等外部知识库中的获取的植物相关知识进行整合，并将获取的全部植物知识存储至 Neo4j 图数据库中，从而实现多源知识融合，其中实体关系与属性类具体知识内容如表 6 所示，具体呈现如图 4 所示。

表 6 先秦典籍植物知识图谱实体及实体关系

编号	实体	实体	关系	数量（组）
实体关系 1	植物	典籍	来源典籍	960
实体关系 2	植物	地点	分布地点	3745
实体关系 3	植物	功效	具有功效	804
实体关系 4	植物	科属	生物类属	315
实体属性 1	植物	学名	具有学名	239
实体属性 2	植物	中文名	具有中文名	315

实体属性 3	植物	别名	具有别名	5680
实体属性 4	植物	其他	具有…属性	881

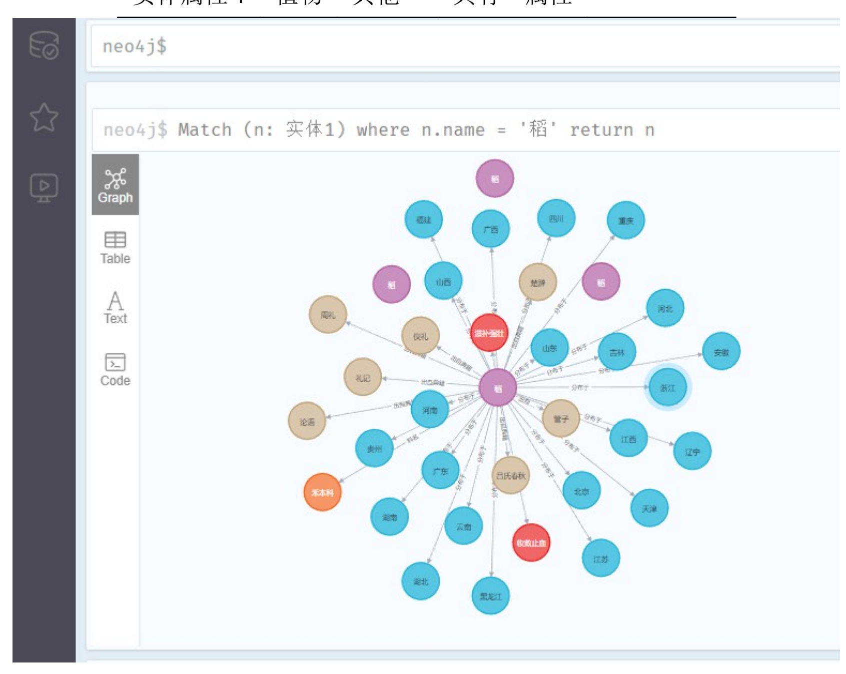


图 4 基于知识图谱的植物关联知识存储结果样例

先秦典籍植物知识可视化呈现

植物与典籍

在图数据库中导入所有典籍名称实体和所有植物实体及其关系后，可以直观观察到两个或者多个典籍共同记载的植物。以《诗经》和《楚辞》为例，如图 5 所示，从该图中可以清晰地看出《诗经》和《楚辞》中记载的相同植物共 26 种，其中包括“桑”“葛”“艾”等植物。此外，特定植物与不同典籍之间的有向边数量也可以一定程度上反映该植物在先秦时期的重要程度。如植物“栗”频频出现于《诗经》《楚辞》等先秦典籍中，而“栗”在当时作为重要的粮作物之一确实占据重要社会地位。同时，根据指向典籍的有向边数量也可以大致估计出该典籍所记载的植物丰富程度。因此，典籍与植物之间的关联关系既为植物研究提供了一个可视化的知识网络，又为文学史研究提供了一个新的有意义的方向。

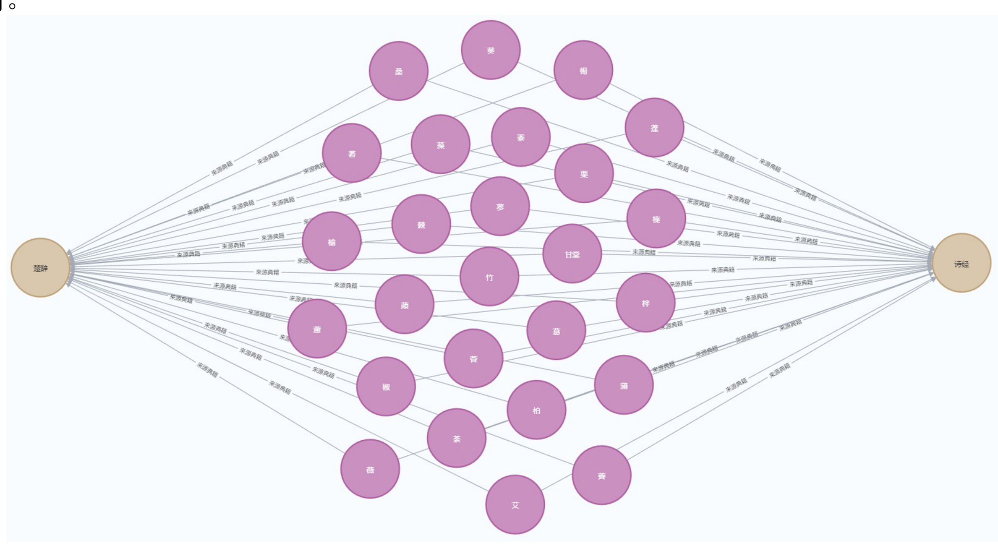


图 5 植物及来源典籍可视化

植物与属性

在图数据库中导入所有典籍植物的所有实体关系和实体属性后，生成的知识图谱如图6所示：这是随机抽取典籍中的一种植物“稻”，对其所包含的知识包括植物的中文名、学名、别称、植物形态特征中的生活型、枝、根、茎、叶、果、花、物候期、生境、海拔国外分布等属性知识进行可视化。可以看出，“稻”有“谷子”“禾”“粳”“稻谷”等别称，是一年生植物，叶鞘松散，无毛，为主要粮食作物之一。此类属性知识的呈现和组织方式拓宽了人们学习植物知识的方法，还为植物研究者提供基于语义的植物知识查询途径，可以根据植物的特定属性信息来限定或者缩小查找范围以便对植物进一步研究。此外除了单个植物的属性知识，先秦典籍植物知识图谱还可以用于探索不同植物的属性之间的关联，比如可以直接查询具有特定属性的植物群进行深入的比较研究，从而挖掘出更具价值的信息。

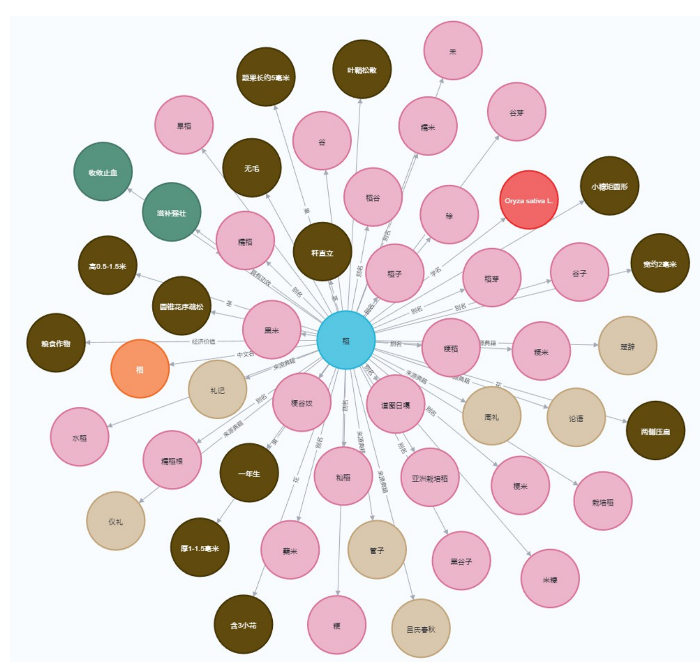


图6 植物属性知识可视化

植物与功效

在先秦时期，植物除了作为中国古代人民粮食作物外，其药用价值也不可忽视。如成书于战国至秦汉时期《皇帝内经》，以及之后的《神农本草经》两部医学著作记载了大量的将植物用药的属性。例如在《本经·序录》中记载的“上药一百二十种为君，主养命以应天，无毒，久服不伤人。”反映了人参、甘草、地黄、黄连、大枣等植物的药用价值。由此可见，植物的部分功效在先秦时期就已经被发现并利用。因此，本研究在构建先秦典籍植物知识图谱过程中分离出植物实体和功效实体，以辅助探究不同植物及其药用价值的关联，部分功效属性可视化如图7所示。

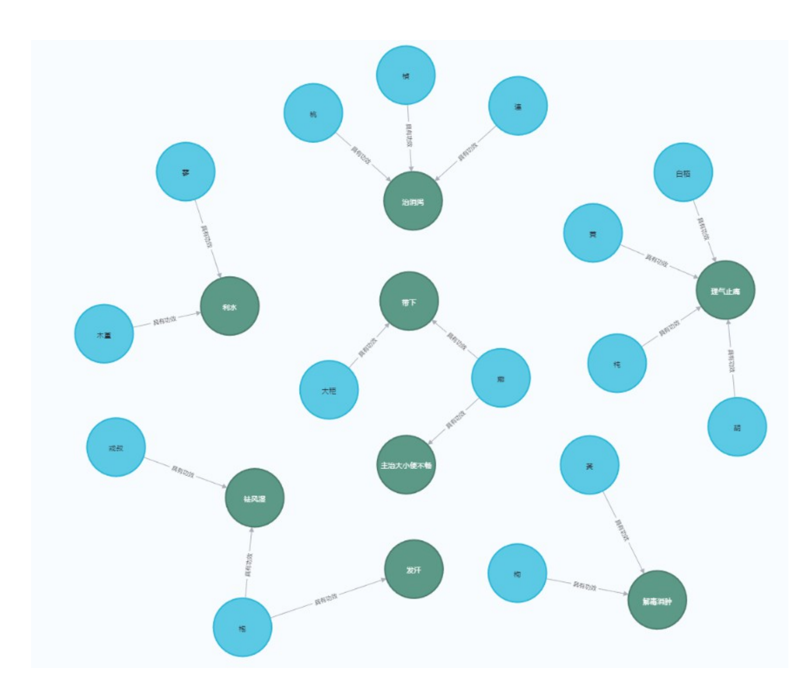


图 7 植物功效知识可视化

图 7 不仅展示了植物具有哪些功效，还对具有相同功效的植物进行关联，如具有“理气止痛”功效的植物有“胡”“杶”等；“戎叔”“枹”等植物均与“祛风湿”相关联。特别是在中药领域，这种展示方式使不同植物与不同功效间的关系更加清晰易懂，不仅有利于研究者发掘具有相似功效的植物，同时也有利于该领域研究者准确查询和了解植物和功效之间的复杂关系，以便更好地研究植物的特性和应用，从而为中药研究提供便利。

结语

本研究对 25 部先秦典籍中的植物词进行了细致地标注，构建了先秦典籍植物实体语料库。基于 CRF、Bi-LSTM-CRF 和多种深度预训练语言模型，构造了面向典籍的古汉语植物实体自动识别模型，为典籍植物知识挖掘提供了有效方法。将从典籍中识别出的植物实体和“植物通”“植物智”等外部百科知识库进行关联整合，构建了先秦典籍植物知识图谱并对知识图谱进行可视化展示。该知识图谱在面向植物的知识发现上具有潜在应用价值，并且可为植物知识检索与自动问答提供数据支撑。

本研究仍存在一些不足之处。首先，古汉语植物实体自动识别模型性能还有进一步提高的空间。例如，“迷穀”、“檗木”等在训练语料中出现频数较低的植物词未能成功识别。其次，先秦典籍植物知识图谱实体、关系、属性等知识仍有待进一步扩充。在知识补全阶段，部分植物关联知识如濒危类别、保护级别、经济价值等尚未完整融合到该图谱中后续研究将考虑增加训练语料的规模，并探索更为先进实体识别方法。此外，在扩充和优化先秦典籍植物知识图谱的同时，还将考虑探索知识图谱在自动问答、知识检索方面的应用。

参考文献

[1] 时唯伟, 周长行, 刘凯, 等. 重金属 Cd 污染土壤的植物修复研究[J]. 中国资源综合利用, 2022, 40(09): 93-95.

[2] 刘冰, 向晓媚, 谭璐, 等. 湖南省德夯峡谷生境种子植物功能性状多样性[J]. 西北植物学报, 2022, 42(9): 1591-1599.

[3] 莫伟军. 药用植物中医保健研究[J]. 核农学报, 2021, 35(03): 768.

[4] 孟迎俊. 《尔雅·释草》名物词研究[D]. 桂林: 广西师范大学, 2010.

[5] 邹俐, 赵焕君, 李娜, 等. 山豆根的本草考证及毒性分析[J]. 现代中医药, 2021, 41(5): 19-23.

[6] 曲保全, 刘墩, 侯文斌, 等. 基于数据挖掘的古代中医典籍外用美发方剂用药规律分析[J]. 中国医药导报, 2021, 18(21): 126-129+149.
[7] 袁代昌, 袁玲, 袁盼盼, 等. 乌药的本草考证[J]. 山西中医, 2021, 37(07): 55-58.
[8] 于娜娜, 王莹莹, 俞静漪. 历史典籍《诗经》中的水生植物和湿生植物意象探析[J]. 湿地科学与管理, 2022, 18(05): 54-57+61.
[9] 谭宏姣. 古汉语植物命名研究[D]. 杭州: 浙江大学, 2004.
[10] 王薇. 《仪礼》名物词研究[D]. 长春: 东北师范大学, 2005.
[11] 王凌云. 中国当代新诗植物意象研究[D]. 昆明: 云南大学, 2017.
[12] 马开颜, 萧瑶, 陈骞, 等. 数字人文视域下中国当代文学作品中的植物意象研究[J]. 数字人文研究, 2022, 2(02): 35-45.
[13] 宋旭晖, 于洪涛, 李邵梅. 基于图注意力网络字词融合的中文命名实体识别[J]. 计算机工程, 2022, 48(10): 298-305.
[14] 李娜, 白振田, 包平. 基于《方志物产》的古籍知识组织路径探析[J]. 古今农业, 2016(01): 105-113.
[15] 徐晨飞, 叶海影, 包平. 基于深度学习的方志物产资料实体自动识别模型构建研究[J]. 数据分析与知识发现, 2020, 4(08): 86-97.
[16] 王菁薇, 肖莉, 骆嘉伟, 等. 基于《伤寒论》的命名实体识别研究[J]. 计算机与数字工程, 2021, 49(08): 1584-1587.
[17] WANG Y, LU L, WU Y, et al. Polymorphic graph attention network for Chinese NER[J]. Expert Systems with Applications, 2022, 203: 117467.
[18] 刘畅, 王东波, 胡昊天, 等. 面向数字人文的融合外部特征的典籍自动分词研究——以SikuBERT 预训练模型为例[J]. 图书馆论坛, 2022, 42(06): 44-54.
[19] WANG Q, MAO Z, WANG B, et al. Knowledge Graph Embedding: A Survey of Approaches and Applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.
[20] 周莉娜, 洪亮, 高子阳. 唐诗知识图谱的构建及其智能知识服务设计[J]. 图书情报工作, 2019, 63(2): 24-33.
[21] 周毅, 刘峥, 粟小青, 等. 融合多层次数据的问答知识图谱本体模型构建[J]. 图书情报工作, 2022, 66(5): 125-132.
[22] 黄微, 卢国强, 赵旭. 基于知识图谱的微博主题演变路径研究[J]. 情报理论与实践, 2022, 45(3): 173-181.
[23] 张君冬. 不孕症中医临床试验知识本体构建研究[D]. 北京: 中国中医科学院, 2022.
[24] 张向先, 李世钰, 沈旺, 等. 数字人文视角下敦煌吐鲁番医药文献知识组织研究[J]. 图书情报工作: 1-16.
[25] 翟东升, 娄莹, 阚慧敏, 等. 基于多源异构数据的中医药知识图谱构建与应用研究[J]. 数据分析与知识发现: 1-23.
[26] 羊艳玲, 李燕, 帅亚琦, 等. 基于中医医案的知识图谱构建[J]. 医学信息学杂志, 2022, 43(10): 50-54.
[27] 李贺, 祝琳琳, 刘嘉宇, 等. 基于本体的简帛医药知识组织研究[J]. 图书情报工作: 1-12.
[28] 崔竞烽, 郑德俊, 王东波, 等. 基于深度学习模型的菊花古典诗词命名实体识别[J]. 情报理论与实践, 2020, 43(11): 150-155.
[29] 张云中, 郭冬, 王亚鸽, 等. 基于知识图谱的红色历史人物知识问答服务框架研究[J]. 图书情报工作, 2021, 65(16): 108-117.
[30] 刘欢, 刘浏, 王东波. 数字人文视角下的领域知识图谱自动问答研究[J]. 科技情报研究, 2022, 4(1): 46-59.
[31] 范青, 史中超, 谈国新. 非物质文化遗产的知识图谱构建[J]. 图书馆论坛, 2021, 41(10): 100-109.
[32] 钟远薪, 夏翠娟. 艺术图像知识图谱构建初探[J]. 图书馆论坛, 2022, 42(02): 109-118.
[33] 张琪, 江川, 纪有书, 等. 面向多领域先秦典籍的分词词性一体化自动标注模型构建[J]. 数据分析与知识发现, 2021, 5(03): 2-11.
[34] 王东波, 刘畅, 朱子赫, 等. SikuBERT 与 SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究[J]. 图书馆论坛, 2022, 42(06): 31-43.

[35] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.
[36] CUI Y, CHE W, LIU T, et al. Pre-Training With Whole Word Masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.
[37] ATTERER M, SCHÜTZE H. Prepositional Phrase Attachment without Oracles[J]. Computational Linguistics, 2007, 33(4): 469-476.

Plant Knowledge Mining and Organization Construction in Pre-Qin Classics from the Perspective of Digital Humanities

Wu Mengcheng [1] Lin Litao [1] Qi Yue[1] Wang Dongbo [1] Liu Liu[1]
[1] College of Information Management, Nanjing Agricultural University, Nanjing 210095

Abstract: [Purpose/significance] The knowledge mining of plants in pre-Qin classics and the construction of pre-Qin plant knowledge map are of great significance for understanding the society and living conditions of ancient Chinese people. **[Method/process]** This paper makes a detailed labeling and quantitative analysis of plant words in pre-Qin classics. Based on CRF and a variety of deep learning models, a plant named entity recognition model for pre-Qin classics was constructed, and the performance of each model was compared and analyzed to determine the optimal model. A knowledge map-oriented knowledge organization model of classics and plants was designed. **[Result/conclusion]** The plant entity recognition model based on the domain pre-trained language model SikuRoBERTa has the best performance, and the harmonic average reaches 85.44%, which provides an effective method for entity-based plant knowledge mining. Aggregation and visualization of plant knowledge in pre-Qin classics.

Keywords: Digital Humanities; Pre-Qin Classics; Plant Named Entity; Deep Learning; Knowledge Graph

作者贡献说明：

吴梦成：确定研究思路和框架，数据标注与分析，论文撰写；
林立涛：研究设计，模型构建，论文修改；
齐月：论文修改；
黄水清：论文修改；
王东波：确定论文选题，论文审阅与修订；
刘浏：论文修改；